A priori reliability of tests with cut score

UNIVERSITÀ DEGLI STUDI DI TORINO

A priori reliability of tests with cut score

Guido Magnano, Chiara Tannoia and Chiara Andrà

Department of Mathematics, University of Turin

Via Carlo Alberto 10, I-10123 Torino (Italy)

guido.magnano@unito.it *

Abstract

The theoretical probability of misclassification in a mastery test is exactly computed using the raw score probability distribution (in the Rasch model) as a function of the examinee's latent ability. The resulting *misclassification probability curve*, together with the latent ability distribution in the group of examinees, completely determines the expected rate of classification errors. It is shown that several distinct ability thresholds, playing different roles in connection to classification reliability, can be associated to a test with a single cut score. In particular, it is possible to define (and compute) two relevant ability

*Part of this work has been done within the TARM Project of the University of Turin. The authors wish to thank prof. Mauro Gasparini for useful discussions, prof. Roger E. Millsap for drawing their attention to Birnbaums original work, and an Associate Editor of *Psychometrika* for suggesting a deeper investigation of the 2PL case.

intervals, which encapsulate the functioning of a mastery test (about and far from the cut score, respectively); the dependence of these intervals on the item difficulty spectrum is investigated. Extension to the 2PL model is also discussed, with emphasis on the effects of weighted scoring.

1 Introduction

This article addresses the problem of how many examinees should be expected to be *incorrectly classified* in a multiple–choice test leading to a pass/fail result (i.e., a *mastery test*). The *pass–fail reliability* (or *classification consistency*) of multiple–choice tests is a classical subject, which has been thoroughly discussed in the sixties by Birnbaum (Lord & Novick, 1968) and subsequently reconsidered, from different viewpoints, in several papers (see e.g. Wilcox (1977), Huynh (1990), Livingston & Lewis (1995), Young & Yoon (1998), Rudner (2005), Wainer *et al.* (2005), Gatti & Buckendahl (2006), Guo (2006) and references therein). Various methods have been proposed to assess the reliability *a posteriori* (i.e. after the test has already been administered, on the basis of the observed results) and/or to define consistency indexes based on asymptotic statistical inference.

The accuracy of a diagnostic test, in general, is measured by the rate of examinees being correctly classified. However, for multiple-choice mastery tests the probability of misclassification is not the same for all individuals:

any overall correct classification rate depends on the ability distribution in the examinees' group, hence such a percentage cannot represent the *intrinsic accuracy* of the test.

(Notice that, in general, *accuracy* would denote the overall rate of correct classifications, while *reliability* would more specifically denote the rate of true positives over all positives: but, in a mastery test, reliability of "fail" and of "pass" results are often equally important, so "pass-fail reliability", "accuracy" and "consistency" are often used as synonyms in this context).

Hence, the starting point is computing the probability of misclassification for a single individual in a given test. This problem presents at least three facets:

i) the definition itself of "misclassification";

ii) the computation of the probability of misclassification for an individual *with a given (latent) ability*;

iii) the computation of the misclassification probability for an individual *who got a specific score in the test.*

Posterior assessment of the overall reliability of a test, i.e. estimates of the total amount of false masters and false non-masters based on observed score distributions, can be obtained from (iii). In contrast, considerations useful for test design should stem from (ii). This article deals with (i) and (ii) only.

The definition of *misclassification* is not quite obvious. Any mastery test involves a pass-fail criterion, which is represented by a *cut score* s_0 (the

minimum score required to pass the test); but to claim that an individual *"should have passed the test, and nevertheless failed"* (or *vice versa*) it is necessary to attribute either a "true score" or a "latent ability" to the examinee. In the true-score approach, it is easy to describe misclassification: an examinee is incorrectly classified in a test if either failed the test although his/her theoretical true score lies above the cut score, or passed the test in spite of having a true score below the cut score (Livingston & Lewis, 1995). Misclassification could also be defined as a mismatch between the results obtained, in a sequence of equivalent tests, by the same group of individuals *(test-retest reliability)*; this, however, entails both theoretical and practical problems (Huynh, 1990).

Within a latent trait model, instead, one should contrast the pass/fail criterion (i.e., the cut score) with a "true mastery" criterion formulated in terms of latent ability. A seemingly natural way to express a mastery criterion is fixing a *threshold ability* (or *mastery level*) θ_0: then, denoting by θ_{lat} the latent ability of a subject and by θ_{obs} the *estimated ability* corresponding to the response to the test, an examinee is misclassified if either $\theta_{obs} < \theta_0 \leq \theta_{lat}$ or $\theta_{lat} \leq \theta_0 < \theta_{obs}$ (Huynh, 1982).

Birnbaum's original setting of the problem in chap. 17 of Lord & Novick (1968), in fact, was different: he assumed that *two* ability thresholds, θ_1 and θ_2, were established by the test makers as "definitely low" and "definitely high", respectively, in connection to the specific purposes of the test. For abilities between θ_1 and θ_2, according to Birnbaum, *"neither classification*

is considered erroneous and no error probabilities are considered". The assumed amplitude of such an "indifference interval" heavily affects the value of any index of classification reliability, since for abilities within this interval both pass and fail results are regarded as correct by definition. Once fixed θ_1 and θ_2, it is possible to seek the item difficulty distribution and the choice of the cut score which would minimize the total misclassification probability. Birnbaum showed that in general this is achieved (in the 2PL model) by choosing items with the highest possible *discriminating power*, and with *difficulties* all belonging to the interval (θ_1, θ_2). In particular, in the limit $\theta_2 \to \theta_1$, all items should have the same difficulty: this distribution is also the one which maximizes the test information function at θ_1.

From Birnbaum's exposition, it may seem that the "indifference interval" could be safely shrunk into a single ability threshold. In fact, the "optimal cut score" is computed using the lower threshold θ_1 (to keep the probability of *false masters* below a given confidence level), while the upper threshold θ_2 is only used to evaluate the probability of *false non-masters*. Yet, Birnbaum shows the plot of the misclassification probability curve for the case of a "large" indifference interval (Lord & Novick, 1968), Fig. 17.4.3, but not the plot which would result while setting $\theta_1 \equiv \theta_2$. It will be shown below that exactly such a plot uncovers the problems arising while considering a single threshold ability.

On the other hand, more recent works on pass-fail reliability make no reference to two ability levels: the authors refer either to a single mastery

ability or to a single cut score. In both cases, the relation binding the pass/fail criterion (i.e., the cut score) to the mastery level remains concealed.

Here it will be assumed that the mastery test aims at separating the population of examinees into exactly *two* ability groups, without devising any "indifference interval": accordingly, a single ability threshold will be set to define "mastery". To describe the relationship between the cut score s_0 and the mastery level θ_0, one should distinguish between two situations:

(A) *the cut score s_0 is set first:* the test givers fixed the cut score upon examination of item content (e.g., using the Angoff procedure), without reference to any specific latent ability level. In this case, to tell whether a subject has been correctly classified or not, one should determine which threshold ability θ_0 corresponds, in some appropriate sense, to the choosen cut score;

(B) *the mastery level θ_0 is set first:* the test is aimed at assessing a predetermined ability threshold, which has been fixed on test-independent grounds (as is assumed in most literature about mastery tests, see e.g. (Huynh, 1980)) and is supposed to remain the same throughout test sessions using different item sets. Then, the question is which cut score s_0 should be adopted, for each test set, to decide whether the sought ability level is reached or not.

In sect. 2 a concrete example, where the different setting in the two cases is illustrated, is used to motivate a number of useful definitions and to state the

main problems: predicting the amount of classification errors, and selecting the test items so to minimize this (expected) amount.

Sect. 3 is devoted to recalling or introducing the definitions and methods which constitute the core of the present investigation. Sect. 4 deals with the problem of relating reliability to the item difficulties' distribution. In Sect. 2–4 only the Rasch model is considered: Sect. 5 concerns the extension of the results to the 2PL model.

Sect. 6 draws the conclusions, which point towards a picture which is somehow reversed with respect to Birnbaum's setting. Instead of starting from an indifference interval (which – once associated with a given confidence level – would determine the choice of the cut score), one can start from a single mastery level: this generates a pair of nested "critical intervals". The largest interval defines a "definitely low" and a "definitely high" ability (in the sense that *outside* that interval the misclassification probability is negligible). The inner interval is instead the neighborhood of the mastery level where the misclassification probability exceeds 0.5. Both intervals can be explicitly calculated if the item difficulties are known. It is shown that only the outer interval is related to the value of the test information function: this explains why, in a number of cases, the common belief that increasing the test information function improves the pass-fail accuracy may be wrong.

2 Setting of the problem

In this section, some general definitions will be given (or recalled) alongside the discussion of a specific example. In this way, the motivations for the definitions themselves and for the subsequent analysis should be clearer.

Suppose that we are dealing with a test set formed by ten items, whose difficulties are assumed to be:

$$\begin{array}{lllll} \beta_1 = -0.89 & \beta_2 = -0.65 & \beta_3 = -0.64 & \beta_4 = -0.58 & \beta_5 = -0.33 \\ \beta_6 = -0.28 & \beta_7 = -0.06 & \beta_8 = 0.04 & \beta_9 = 0.35 & \beta_{10} = 0.81 \end{array} \tag{1}$$

As anticipated in the introduction, the same test set can be used in two different ways. In case (A) the test makers decided that the cut score s_0 should be set, for instance, to 6. In case (B), instead, test makers intend to use the test set (1) to assess a predetermined mastery level, say $\theta_0 = -0.20$.

2.1 Case A: predetermined cut score

In our example, the cut score is *ab initio* set to 6. Then, one has to identify the ability level which should, in principle, lead to a "correct" pass result.

For a test with N items, it is possible to compute the average *(number-right)* raw score expected for a given latent ability. A correspondence between the score scale and the abiity scale is then given, in the Rasch model, as follows (Baker, 1992):

Definition 1 *The* Test Characteristic Curve *maps any ability θ to the cor-*

responding expected value $S(\theta)$ of the number-right score:

$$S(\theta) = \sum_{i=1}^{N} P_i(\theta), \tag{2}$$

where $P_i(\theta)$ is the probability of correct response to the i-th item. In the Rasch model, where

$$P_i(\theta) = \frac{1}{1 + e^{(\beta_i - \theta)}}, \tag{3}$$

the maximum likelihood estimate for the ability is the (unique) preimage of the observed raw score under the TCC; we shall denote this preimage by $\Theta(s)$.

Neither of the extremal scores, $s = 0$ and $s = N$, corresponds to a finite ability estimate; in common practice it is customary to assign two conventional finite ability values to extremal scores, but this is irrelevant for the present purposes. It is important, instead, to keep in mind that – since only integer scores can be obtained – estimated abilities can take only a discrete set of values:

Definition 2 *In a given Rasch test, the only* observable abilities*, i.e. the possible ability ML estimates, are the TCC preimages $\Theta(n)$ of the integer scores $n = 1 \ldots (N-1)$.*

The distance between two consecutive observable abilities is determined by the slope of the TCC. The latter is given by the sum of the derivatives of the probabilities (3): each of these derivatives is maximal at $\theta = \beta_i$. Therefore, the maximum possible slope of a TCC, for a test with N items,

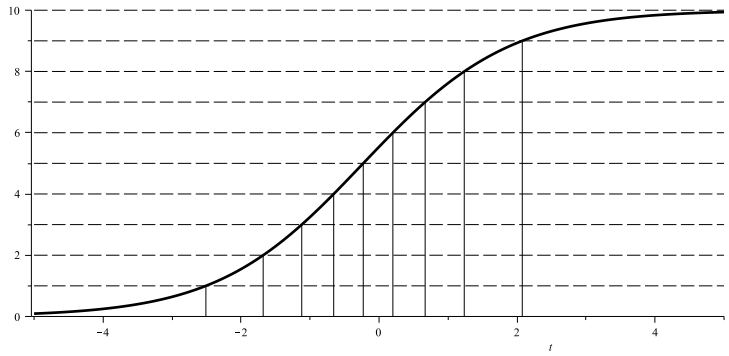


Figure 1: TCC *(Test Characteristic Curve)* for the sample test (1)

is attained if all β_i are equal to some given value β, is located at $\theta = \beta$ and equals $\frac{N}{4}$, as can be easily checked. The slope at any other point (and at any point, in the case of a generic item difficulty spectrum) is always lower. It follows that

Proposition 1 *In a Rasch test with N items, the distance between two observable abilities is always greater than $\frac{4}{N}$ (in the logit scale).*

Intermediate ability values have no chances at all of being attributed to an examinee when using the given test set. This contrasts, for tests with few items, with the widespread assumption that the ability estimate be normally distributed around the true latent ability (this is, instead, the asymptotic limit for large N).

For the sample test (1), Fig.1 shows the plot of the TCC, and here is the list of observable abilities $\Theta(s)$ (rounded to the second decimal digit), which correspond to the vertical lines in the figure:

$$\begin{array}{lll} \Theta(1) = -2.51 & \Theta(4) = -0.66 & \Theta(7) = 0.67 \\ \Theta(2) = -1.68 & \Theta(5) = -0.23 & \Theta(8) = 1.23 \\ \Theta(3) = -1.12 & \Theta(6) = 0.20 & \Theta(9) = 2.08 \end{array} \tag{4}$$

The ability corresponding to the cut score $s_0 = 6$ under the TCC is thus $\Theta(s_0) = 0.20$. It might seem natural to regard this ability as the mastery level associated to the cut score. With that threshold, an individual having ability *below 0.20* and *passing the test* should be considered a "false master", while individuals with ability greater or equal to 0.20 and failing the test would be regarded as "false non-master".

On this basis, the theoretical misclassification probability for a given latent ability can be exactly computed, as will be shown later. It turns out that for an ability of 0.10, for instance, the probability of passing the test (resulting a false master) is 0.58: such an individual has therefore *higher chances of being classified incorrectly than correctly.*

It may seem surprising that the error probability can exceed 0.5. To ensure that the error probability is nowhere greater than 0.5 one should, in fact, consider a *different* ability threshold: namely, the ability level such that the probability of passing the test exactly equals 0.5.

Definition 3 *Given a test set and a cut score, the* critical ability*, hereby denoted by θ_c, is the latent ability giving equal probabilities of passing or failing the test.*

As shown in the next section, the ability θ_c is uniquely defined in this way.

For the sample test (1) with cut score at 6, the critical ability turns out to be $\theta_c = -0.024$, much lower than the threshold previously considered.

Choosing θ_c instead of $\Theta(s_0)$ as the "mastery level" *changes* the definition of misclassification for all individuals falling within the ability interval between these two values. For test set (1), for instance, the chances of misclassification for an individual with latent ability 0.18 would drop from 64% (the probability of passing the test) to 36% (the probability of failing the test).

Adopting the critical ability as the mastery level for a test with predetermined cut score solves the problem of ensuring that the individual misclassification probabilities never exceed 0.5. Any other choice of the mastery threshold will necessarily raise the misclassification probability above 0.5 in some range of abilities.

2.2 Case B: predetermined mastery level

In contrast with the previous case, a mastery level $\theta_0 = -0.20$ has been assumed *a priori*. Then, two questions arise: first, which is the appropriate cut score?

A Rasch ML estimate would attribute an ability level $\theta_{obs} = -0.23$ to any individual scoring 5, and $\theta_{obs} = 0.20$ to any individual scoring 6, as shown in (4). Hence, requiring $\theta_{obs} \geq \theta_0$ entails setting the cut score equal to 6. More generally, in case (B) the cut score for the test set is uniquely determined by the mastery level:

Proposition 2 *In a given Rasch test, if MLE is used to connect the ability to the observed score, assuming a threshold ability level θ_0 necessarily sets the cut score s_0 to the lowest integer greater than (or equal to) $S(\theta_0)$.*

The next question then becomes: is the test set really apt to assess the mastery level $\theta_0 = -0.20$? As already mentioned, Birnbaum's analysis shows that in an optimal mastery test set all items should have the same difficuly. But in real situations one is unable to produce at will items having a predetermined difficulty level. Test makers, in the best possible situation, have at their disposal a large set of items whose difficulty is known with good accuracy after calibration in previous tests, and can select the test items from that item pool. The problem is thus selecting a "good" test set, hopefully the best one, *among all concretely available choices*.

Now, the average difficulty of the test set (1) is $\bar{\beta} = -0.223$; the standard measurement error (the square root of the inverse of the test information function) at θ_0 turns out to be $\sigma = 0.65$, while for an "ideal" test with 10 items with equal difficulty $\beta = \theta_0$ the standard error would be $\sigma = 0.63$. Thus, judging from the fact that the average difficulty of the test is fairly close to θ_0 (only 0.02 logit higher), and that the expected measurement error is not far from the "optimal" test, the choice of test set (1) would hardly seem unreasonable. And yet, it will be shown below that this conclusion is fallacious: in order to foresee the reliability of the test one should rely on a different analysis, which is the subject of the next sections.

It should now be clear that the situations depicted as A and B are differ-

ent, *although the test set and the cut score s_0 are exactly the same*, because the definition of misclassification depends on whether one assumes the mastery level θ_0, in our example, to be equal to $\Theta(6)$, to θ_c or to -0.20 (any threshold ability greater than -0.23 and not exceeding 0.20 would in fact yield $s_0 = 6$). The differences in misclassification probabilities only affect latent abilities in the range between two consecutive observable abilities, in our case $\Theta(5)$ and $\Theta(6)$. The disparity would thus disappear upon adopting Birnbaum's setting of *two* mastery thresholds, provided these are separated by an "indifference interval" large enough to contain both $\Theta(s_0 - 1)$ and $\Theta(s_0)$.

2.3 Evaluating the misclassification probability

Once chosen an appropriate misclassification criterion, theoretical computation of the probability that an individual with a given ability is incorrectly classified presents no ambiguity. This is discussed in the next sections; here, instead, the *misclassification frequencies obtained in a computer simulation* for the test set (1) will be confronted with probabilities estimated according to the method introduced in (Rudner, 2005).

According to this method, the approximate probability of incorrect classification would be obtained as follows. The ability estimate θ_{obs} for an examinee with given latent ability is assumed to be normally distributed, whereby the mean of the distribution coincides with the latent ability, $\mu = \theta_{lat}$, and the variance equals to the inverse of the *test information function* evaluated

θ_{lat}	$\theta_0 = -0.024$		$\theta_0 = 0.20$		$\theta_0 = -0.20$	
	% obs.	% exp.	% obs.	% exp.	% obs.	% exp.
-0.723	13.6	14.7	13.6	8.30	13.6	21.6
-0.612	17.9	18.6	17.9	10.9	17.9	26.6
-0.501	21.8	23.3	21.8	14.2	21.8	32.3
-0.390	28.4	28.7	28.4	18.2	28.4	38.5
-0.279	34.4	34.7	34.4	23.0	34.4	45.2
-0.167	40.9	41.3	40.9	28.6	**59.1**	48.0
-0.056	47.7	48.0	47.7	34.8	**52.3**	41.3
0.055	44.9	45.2	**55.1**	41.2	44.9	34.8
0.166	37.8	38.7	**62.2**	47.9	37.8	29.0
0.277	31.4	32.7	31.4	45.4	31.4	23.9

Table 1: Misclassification rates (observed vs. expected according to Rudner's estimate) for test set (1), for different mastery levels compatible with the cut score $s_0 = 6$.

at the latent ability, $\sigma^2 = F(\theta_{lat})^{-1}$. Then, the probabilty of misclassification P_m for the ability θ_{lat} is given by the value of the corresponding normal cumulative distribution $\Phi_{\mu,\sigma}$ at the mastery threshold. Namely, $P_m = \Phi_{\mu,\sigma}(\theta_0)$ if $\theta_{lat} \geq \theta_0$, and $P_m = 1 - \Phi_{\mu,\sigma}(\theta_0)$ if $\theta_{lat} < \theta_0$. The highest misclassification probability obtained in this way is always 0.5 and occurs when the latent ability exactly equals the mastery level, independently of the test set. In Table 1, misclassification rates observed in computer simulation (using a sample of 10 000 "virtual individuals" for each ability level) and (Rudner) expected rates are compared, for the three possible mastery levels considered in the previous discussion. Ten reference abilities have been chosen at equal distances, centered on the average item difficulty of the test.

While the expected rates reasonably match the observed rates when θ_0 is set equal to the critical ability $\theta_c = -0.024$, the disagreement is noticeable

in the other two cases. Rates of misclassification higher than 50% cannot be expected within Rudner's approach, and nevertheless they occur. Even outside the ability range $(-0.20, 0.20)$, the discrepancy between observed and expected rates can exceed 10%. It can be seen that:

- in situation A, i.e. when the mastery level can be chosen to suit the given test set and cut score, setting it equal to the critical ability θ_c produces the lowest misclassification rates; with this choice, moreover, Rudner's estimates of such rates turn out to be fairly accurate. In contrast, regarding $\Theta(s_0)$ as the mastery level for this test leads to large misclassification rates (underestimated by Rudner's method);

- for situation B, observed error rates reveal that test set (1) is actually *unsuited* to assess the sought mastery level $\theta_0 = -0.20$. This fact could not be detected from the average difficulty of the test set, nor from the value of the test information function. Actually, it does not even depend on the gap between θ_0 and $\Theta(6)$: as will be shown in the next section, it is instead the distance between θ_0 and θ_c which matters.

The mere observation that the assumed mastery level $\theta_0 = -0.20$ is very much closer to $\Theta(5)$ than to $\Theta(6)$ might lead one to guess that, in situation B, lowering the cut score to 5 would considerably reduce misclassification. On the contrary, the overall situation would not improve at all: only, the highest misclassification rates would be shifted to lower abilities. In fact, with a cut score of 5 and with the same mastery level $\theta_0 = -0.20$, the ob-

served misclassification rates for latent abilities -0.390 and -0.279 become, respectively, 53.5% and 59.4%. Moreover, fixing the cut score at 5 would be inconsistent with the fact that $\Theta(5) < \theta_0$.

In conclusion, in situation A a viable strategy exists in order to minimize, compatibly with the given test set, both misclassification and error estimation issues: the strategy consists in choosing the critical ability as the mastery level corresponding to the given cut score.

In situation B, instead, nothing can be done but changing the test set itself to match the mastery level, but it is not evident how to do so. Connecting misclassification with the standard measurement error would suggest that the test set should be chosen so to maximize the information function at the examinee's true ability, which may be endeavored using adaptive testing (adaptive mastery tests are indeed a particular case of situation B). On the other hand, it has been proven by Birnbaum that for an "optimal" mastery test the information function should reach its maximum at the mastery level, not at the examinee's ability: this seems to be an argument against the use of adaptive tests for mastery assessment.

But the example discussed so far shows that the value of the test information function at the mastery level does not reveal, by itself, the extent of the reliability issues: what is more, increasing the information function does not always improve the pass-fail reliability. Going back to the example, suppose that test makers manage to substitute both the easiest and the most difficult item in (1) with two new items having difficulty exactly equal to $\theta_0 = -0.20$.

It can be seen (by computing the TTC) that for the new test set the cut score is still 6, and the test information function at the threshold is indeed increased (from 2.361 to 2.443, against a maximum possible value of 2.5 for a 10-item test set). Performing again a computer simulation (with the same reference abilities) for the new test set, one finds that the highest misclassification probability in the table (at $\theta_{lat} = -0.167$) is reduced from 59.1% to 57.8%, and the misclassification probabilities for higher abilities (false non-masters) are reduced as well, but at the price of raising all the error probabilities for lower abilities (false masters): for instance, for $\theta_{lat} = -0.390$ the misclassification probability is raised from 28.4% to 30.8%. Hence, depending on the ability distribution in the population, the "improved" test might actually produce a *larger* amount of misclassifieds.

One would expect such phenomena to fade away for test sets including a larger number of items. This is only partly true: misclassification, in the range between the critical ability for the test set and the assumed mastery level, will always prevail over correct classification. Thus, further theoretical insight is needed.

3 Misclassification Probability Curve

A recursive procedure to compute exactly the probability $P(s|\theta)$ that an examinee with latent ability θ obtain the raw score s can be found in (Lord & Novick, 1968) or in (Lord & Wingersky, 1984). An alternative method is

presented in (Khidr & Abdelnasser, 1982). Summing these probabilities for all scores $s \geq s_0$ gives the probability that the examinee pass the test.

$$P_{pass}(\theta; s_0) = \sum_{k=s_0}^{N} P(k|\theta) \qquad P_{fail}(\theta; s_0) = \sum_{k=0}^{s_0-1} P(k|\theta) = 1 - P_{pass}(\theta; s_0) \quad (5)$$

Definition 4 *The* Pass Probability Curve $P_{pass}(\theta; s_0)$ *gives the theoretical probability of getting a score at least equal to* s_0*, as a function of the latent ability of the examinee.*

(Note: as previously recalled, the score probability distribution $P(s|\theta)$ can be straightforwardly translated – through the TCC – into a probability distribution for the estimated ability. The latter distribution is known to be asymptotically normal for large N, but normality is nowhere assumed in the following computations)

The PPC is a close relative to the PPoP curve introduced by Wainer (Wainer *et al.* , 2005). There is however a conceptual difference: the PPoP curve is defined to be a posterior probability curve constructed from the *observed scores*, while the PPC is theoretically derived from the assumptions of the Rasch model and from the knowledge of the item difficulties.

The curve P_{pass} is always monotone, and so is P_{fail}: the two curves intersect only at a single critical ability θ_c. In a sense, for a dichotomous (pass/fail) test the PPC (not the TCC) plays the same role as the Item Characteristic Curve for a single item; similarly to 2PL model (although the PPC is not, in general, a logistic curve), θ_c plays the role of the difficulty parameter β, and

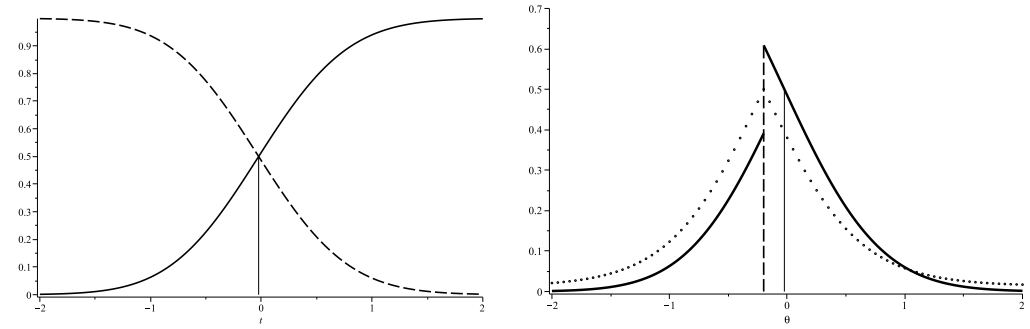
the slope of the PPC at θ_c somehow represents the *discriminating power* of the mastery test. Analogous remarks can be found in (Lord & Novick, 1968) (p. 409) and in (Wainer *et al.* , 2005). A notable fact, that has already been used in the previous section, is the following:

Proposition 3 *For a Rasch test with cut score s_0, the critical ability belongs to the interval between the observable abilities $\Theta(s_0 - 1)$ and $\Theta(s_0)$:*

$$\Theta(s_0 - 1) < \theta_c \leq \Theta(s_0). \tag{6}$$

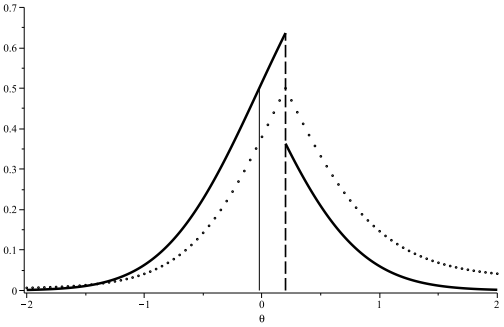
In fact, when $s(\theta)$ (the expected score) is an integer number, then $s(\theta)$ is both the mean and the mode of the score probability distribution for the ability θ. From the mean-median-mode inequality and from the definition of median for discrete distributions, it follows that for the ability θ the probabilities that the score is strictly less than $s(\theta)$ and that the score is strictly greater than $s(\theta)$ are both less than 0.5. Applying this fact to the abilities $\Theta(s_0 - 1)$ and $\Theta(s_0)$, one finds that $P_{pass}(\Theta(s_0 - 1); s_0) < 0.5$ and $P_{pass}(\Theta(s_0); s_0) \geq 0.5$, which proves the proposition.

Now, suppose that a mastery level θ_0 has somehow been established, not necessarily equal to θ_c. The probability of incorrect classification equals $P_{pass}(\theta; s_0)$ for $\theta < \theta_0$ and $P_{fail}(\theta; s_0)$ for $\theta \geq \theta_0$. The plot of these probabilities will be called the *misclassification probability curve (MPC)*. Fig. 2 (b, c, d) show the MPC (solid curve) for the test set (1) with $s_0 = 6$, under the three different values of θ_0 considered in the previous section. The graphs

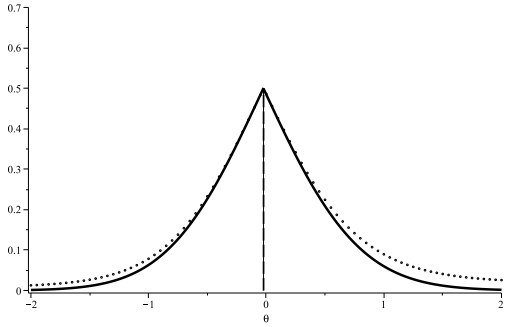


(a) P_{pass} and P_{fail} for test set (1)

(b) MPC for $\theta_0 = -0.20$



(c) MPC for $\theta_0 = 0.20$



(d) MPC for $\theta_0 = \theta_c = -0.024$

Figure 2: PPC and Misclassification Probability Curves for test set (1). The location of the critical ability θ_c is represented by a solid vertical line, while the dashed vertical lines correspond to the assumed mastery levels θ_0. Dotted curves represent Rudner's approximation of MPC.

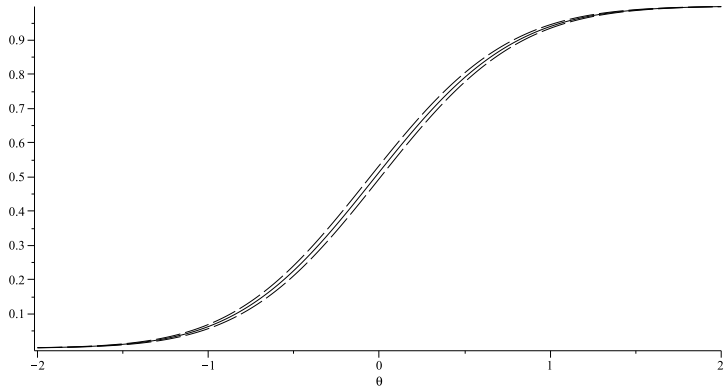


Figure 3: PPC standard deviation range for the sample test (1), assuming a measurement error of 0.1 logit on the item difficulties.

clarify why for examinees ranging in the interval (θ_0, θ_c) the chances of incorrect classification are always greater than 50%, and why setting $\theta_0 = \theta_c$ is the only way to ensure that misclassification probabilities never exceed 0.5. The values of the computer-simulated frequencies listed in the previous section agree with the probability values computed theoretically, and the source of the discrepancies with the approximate values given by Rudner's method when $\theta_0 \neq \theta_c$ becomes evident.

3.1 Effect of item difficulty uncertainty

Throughout this article it is always assumed that the item difficulties are exactly known. This assumption is common in reliability analysis, as well as when dealing with adaptive test construction. To judge to what extent the PPC of the sample test set (1) would be affected by the uncertainty on the item parameters, assume e.g. a standard measurement error of 0.1 logit for the items difficulty (roughly, this would be obtained after testing the items on 400 individuals). Fig. 3 displays the corresponding variability of the PPC curve (as obtained by adding to each β_i a "gaussian noise" with $\sigma = 0.1$). One should be warned, however, that both the cut score and the critical ability undergo large oscillations if $\Theta(\theta_0)$ is close to an integer value.

4 Reliability and item difficulty spectrum

4.1 Computational evidence

It is reasonable, at this point, to ask to what extent the above considerations could help to design a test with optimal pass/fail reliability. To disentangle the complex interplay between the various relevant factors (including the population ability distribution) it is useful to look at the examples included in Tab. 2, produced in the following way.

The mastery ability (determining the cut score) is now assumed to be $\theta_0 = 0$; several test sets have been considered, and a number of reliability–connected parameters have been computed (and checked against simulated administration of the tests) for two simulated populations. The first one (population A) is a sample of 1000 ability levels (virtual individuals) taken from a standard normal distribution ($\mu = 0, \sigma = 1$), so that masters and non-masters are in almost equal number. Population B is instead a sample of 1000 ability levels taken from a normal distribution with $\mu = 1$ and $\sigma = 1$, so that true non-masters are only 15% of the population. The test sets have been constructed by varying the number of items, the average difficulty and the difficulty range of the items, in all possible combinations within the following scheme:

- The number of items, N, has been set to be either 11, 21, 31 or 41 (the reason for choosing odd numbers is explained in §4.2).
- The average difficulty of the items has ben set equal to either 0, 1 or -1;

sets with average difficulty equal to 1 have not been used for population A, because the results would be substantially symmetric with respect to sets with average difficulty equal to -1. For population B, two further sets with $N = 11$ and average difficulty at 0.5 have been considered.

- The items of each set have been chosen either to have all the same difficulty, or to have a spectrum of difficulties "equidistributed" (i.e. spread at equal distances) in a range of ± 2 logits. Both cases are purely fictional, but can be regarded as the limit cases for highly concentrated or "wide rectangular" difficulty distributions, respectively.

Table 2 displays the following data:

Columns 4 to 8 contain values which are intrinsic to the set itself, not involving the population to which the set is administered: the cut score, the values of the Test Information Function at θ_0 and at θ_c, the *critical ability* and the *critical interval amplitude*. The latter is the width of the ability range for which the misclassification probability exceeds 0.1: the amplitude numerically computed from the MPC is followed, in parentheses, by the approximate amplitude computed using the formula (9) given in sect. 4.3 below.

Columns 9 and 12 show the value of the *expected rate of misclassifieds*, for population A and B respectively.

Columns 10 and 13 display the number of expected false non-masters over the expected number of examinees failing the test, in each population; the same for false masters in columns 11 and 14. All these values have been

Table 2 - Comparison of 26 different test sets for the mastery level $\vartheta_0 = 0$, for two populations A e B

set id	# of items	difficulty distribution	cut score	IF(0)	IF(ϑ_c)	critical ability	α = 0.1 critical interval amplitude	pop. A: true non-masters = 487, true masters = 513			pop. B: true non-masters = 153, true masters = 847		
								% misclass.	false nonmast. / exp. fail	false mast. / exp. pass	% misclass.	false nonmast. / exp. fail	false mast. / exp. pass
1	**11**	all β = 0	6	2.75	2.75	0.00	1.53 (1.55)	16.5	85 / 492	80 / 508	11.3	75 / 189	38 / 811
2		equid. in [-2, 2]	6	2.00	2.00	0.00	1.78 (1.81)	18.7	96 / 492	91 / 508	13.0	89 / 200	41 / 800
3		all β = -1	9	2.16	1.98	0.18	1.80 (1.82)	19.3	132 / 558	61 / 442	16.0	130 / 254	30 / 746
4		equid. in [-3, 1]	8	1.83	1.82	0.01	1.86 (1.90)	19.4	104 / 500	90 / 500	14.0	99 / 211	41 / 789
5		all β = 1	3	2.16	1.98	-0.18	1.80 (1.82)				11.1	56 / 155	54 / 845
6		equid. in [-1, 3]	4	1.83	1.82	-0.01	1.86 (1.90)				13.2	88 / 198	44 / 802
7		all β = 0.5	5	2.59	2.66	0.14	1.55 (1.57)				13.0	100 / 224	30 / 776
8		equid. in [-1.5, 2.5]	5	1.96	1.96	0.00	1.80 (1.83)				13.1	89 / 201	41 / 799
9	**21**	all β = 0	11	5.25	5.25	0.00	1.11 (1.12)	12.3	63 / 491	59 / 509	8.3	51 / 172	32 / 828
10		equid. in [-2, 2]	11	3.91	3.91	0.00	1.28 (1.30)	14.1	73 / 491	68 / 509	9.5	60 / 179	35 / 821
11		all β = -1	16	4.13	4.10	0.02	1.26 (1.27)	13.8	77 / 503	61 / 497	9.8	65 / 186	32 / 814
12		equid. in [-3, 1]	15	3.54	3.50	0.05	1.36 (1.37)	14.9	87 / 512	62 / 488	10.7	75 / 196	32 / 804
13		all β = 1	6	4.13	4.10	-0.02	1.26 (1.27)				8.9	53 / 170	36 / 830
14		equid. in [-1, 3]	7	3.54	3.50	-0.05	1.36 (1.37)				9.5	55 / 168	40 / 832
15	**31**	all β = 0	16	7.75	7.75	0.00	0.91 (0.92)	10.2	52 / 490	49 / 509	6.9	40 / 165	28 / 835
16		equid. in [-2, 2]	16	5.82	5.82	0.00	1.05 (1.06)	11.7	60 / 491	57 / 509	7.9	48 / 170	31 / 830
17		all β = -1	23	6.09	6.20	0.04	1.02 (1.03)	11.4	54 / 481	60 / 519	7.6	43 / 163	33 / 837
18		equid. in [-3, 1]	22	5.25	5.17	0.06	1.12 (1.13)	12.6	77 / 516	48 / 484	9.0	63 / 190	27 / 810
19		all β = 1	9	6.09	6.20	0.04	1.02 (1.03)				7.8	50 / 175	28 / 825
20		equid. in [-1, 3]	10	5.25	5.17	0.06	1.12 (1.13)				7.8	41 / 156	38 / 844
21	**41**	all β = 0	21	10.25	10.25	0.00	0.80 (0.80)	8.8	46 / 490	43 / 510	6.0	34 / 162	26 / 838
22		equid. in [-2, 2]	21	7.72	7.72	0.00	0.92 (0.92)	10.2	52 / 490	49 / 510	6.9	40 / 165	28 / 835
23		all β = -1	30	8.06	8.31	0.07	0.89 (0.89)	10.0	41 / 469	59 / 531	6.5	31 / 151	33 / 849
24		equid. in [-3, 1]	28	6.96	7.06	0.07	0.96 (0.96)	10.8	44 / 467	64 / 533	6.9	34 / 152	35 / 848
25		all β = 1	12	8.06	8.31	0.07	0.89 (0.89)				7.0	48 / 178	22 / 822
26		equid. in [-1, 3]	14	6.96	7.06	0.07	0.96 (0.96)				7.7	53 / 183	23 / 817

computed using the MPC; these rates, as well as the total misclassification rates, are in full agreement with the mean values (not reported) that we have observed in 50 computer simulations for each test set and for each population.

A number of facts come out from table 2. The overall misclassification incidence (col. 9 and 12) depends, as expected, on both the test set and the ability distribution in the population. For each test, the incidence is always lower for population B. The reason is that the highest misclassification probability is encountered near to the mastery level: the latter is close to the population mode for population A, while for population B it falls in left tail of the distribution.

For each population, the overall misclassification rate primarily depends on the number of items in the test. However, there are some exceptions to the rule "the higher the information function, the higher the reliability": relevant exceptions are encountered when the critical ability is significantly different from zero (e.g. set 4/B vs. set 3/B).

The overall misclassification rate, on the other hand, is not necessarily the most relevant issue, for the cost of misclassification may be different for false positives and false negatives (van der Linden, 1998). Tests with similar overall accuracy may behave in quite different ways as far as the reliablity of false positives (or of false negatives) is concerned. For population A, test set 1 of table 2 will produce approximately the same number of false positives and false negatives, while test set 3 yields twice more false non-masters than false masters. The rates of false positives and false negatives

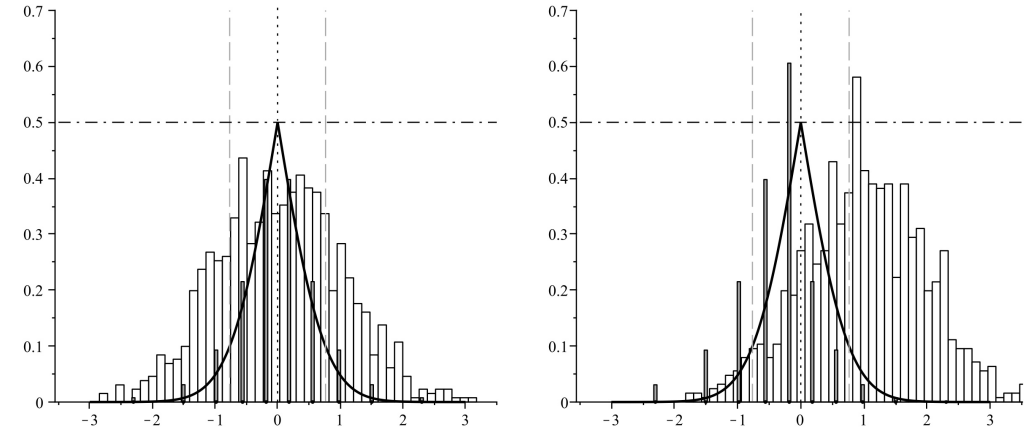
strongly differ whenever the population ability distribution is asymmetric with respect to the mastery ability, and/or whenever the test critical ability θ_c differs significantly from θ_0. When the two effects have opposite signs (set 5/B) they partially cancel each other.

The *relative* incidence of false positives or false negatives is also worth considering. On population B, the test set 1 has an expected overall error rate of 11.3%: yet, an individual who *fails* such a test has a chance of almost 40% of having been incorrectly classified. For the test set 3, this probability raises to 51%: even if the overall error rate is 16%, examinees *failing* this test are incorrectly classified in the majority of cases!

As stressed in the introduction, the *a priori* misclassification probability for a given latent ability should not be confused with the *posterior* misclassification probability for an examinee who got a given score in the test. The latter will be discussed elsewhere, but a comparison is shown in fig. 4, where the dark vertical columns show the relative frequency of misclassifieds among examinees who obtained a given score (the columns are located at the corresponding estimated abilities): such posterior misclassification rates (which can exceed 50% as well) depend on the population ability distribution, which is depicted by the white histogram behind each plot.

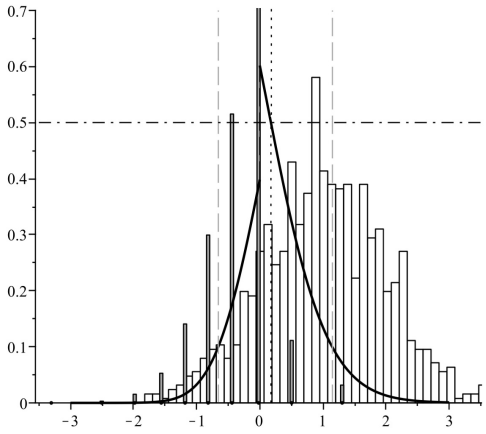
4.2 "Optimal" test sets: odd and even N

It has been shown so far that the *ideally* optimal mastery test (for a given number N of items) should have both the highest possible PPC slope at θ_c

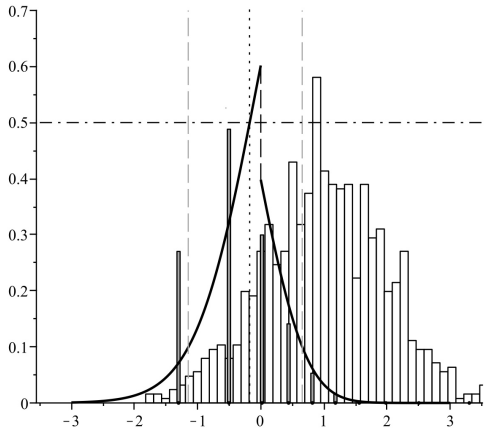


(a) Test set 1, pop. A

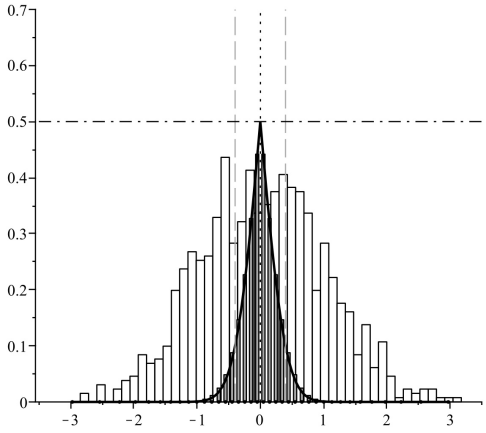
(b) Test set 1, pop. B



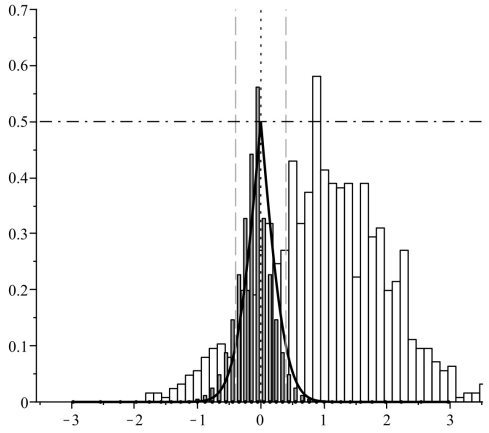
(c) Test set 3, pop. B



(d) Test set 5, pop. B

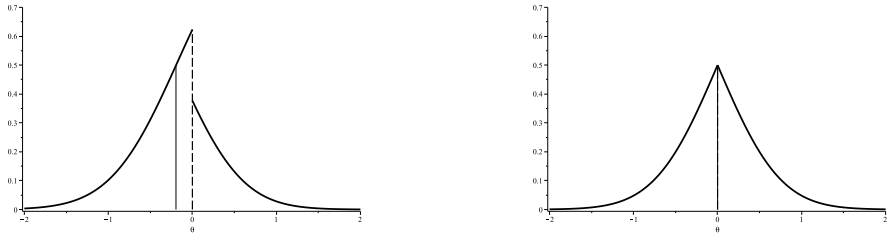


(e) Test set 21, pop. A



(f) Test set 21, pop. B

Figure 4: MPC curves, critical ability (vertical dotted line) and (0.1) critical interval (vertical dashed lines) for some of the test sets of table 2. The white histograms show the population ability distributions; the grey columns represent the *posterior* misclassification probability for a given observed score.



(a) **(b)**

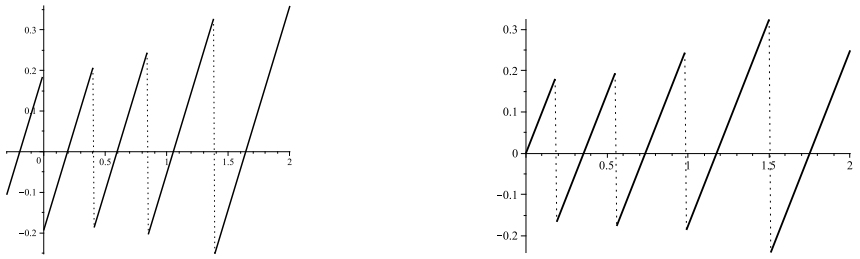
Figure 5: MPC for "ideal" test sets with difficulties concentrated at $\beta = \theta_0 = 0$. **(a)**: test set with 10 items; **(b)**: test set with 11 items

and the critical ability coincident with the assumed mastery level, $\theta_c = \theta_0$. If N is odd, both requirements are met (on Rasch assumptions) if all items have difficulty equal to θ_0: then, the pass and fail probability curves are mutually symmetric with respect to θ_0, and θ_c lies exactly at that point (fig. 5b).

When N is even, instead, if $\beta = \theta_0$ for all items then the cut score will be $s_0 = \frac{N}{2}$. At $\theta = \beta$ the probability of passing the test is then larger than 0.5, hence the critical ability is forcefully lower than β (fig. 5a). Therefore, with an even number of items the difficulty should not be centered at the sought mastery level θ_0, but at the slightly higher value (numerically computable) such that $\theta_c = \theta_0$. More generally, for N even, $\theta_c \neq \theta_0$ if items are symmetrically distributed around θ_0. In summary,

Proposition 4 *For a given mastery ability θ_0 and a given number of items N, a test set where all items have difficulty exactly equal to θ_0 is optimal (in the sense that the PPC slope is maximal and $\theta_c = \theta_0$) if and only if N is odd.*

The dependence of the critical ability on item parameters is complicated, and actually discontinuous, as can be seen in fig. 6. Discontinuities occur



(a) **(b)**

Figure 6: Discontinuos variation of the critical ability for test sets concentrated at a single difficulty β. The mastery level is kept fixed, $\theta_0 = 0$, while β varies. The critical ability is undefined at discontinuity points, which correspond to jumps of the cut score. **(a)**: test set with 10 items; **(b)**: test set with 11 items. With $N = 10$, to get $\theta_c = \theta_0$, one should set $\beta \neq 0$.

whenever the mastery level corresponds (through the TCC) to an integer expected score. About such points, a minimal change of the item distribution causes the cut score to jump up to the next value, and both the PPC and the critical ability shift abruptly: hence the cut score itself, the critical ability and the MPC curve all become highly unstable with respect to measurement uncertainty. This is why only ideal tests with odd N have been considered in table 2: for ideal test sets with an even number of items the corresponding values would be unstable, and therefore not representative of real tests' behavior.

4.3 Critical misclassification interval

In table 2 one can observe a high correlation between the overall error rate and the width of the interval defined as follows:

Definition 5 *The* critical (misclassification) interval *for the probability level α is the ability range for which the value of the misclassification probability (MPC) is greater than or equal to α.*

In other terms, any individual with ability belonging to the critical interval has a probability at least equal to α of being misclassified. The critical amplitude, the PPC slope and the information function value at the critical ability are strongly correlated to each other: the latter is much easier to compute, but the critical interval provides more direct information on *which portion of the population* has significant chances of being incorrectly classified.

Computing the critical interval for a given test amounts to finding at which points the PPC equals α and $1-\alpha$, respectively. In general, this should be done numerically. The minimal critical amplitude for a test with N items corresponds to the highest possible slope of the PPC at the critical ability, which occurs (for N odd) for ideal tests with all items concentrated at the mastery ability. For ideally "concentrated" tests with $\beta = \theta_0 = \theta_c$ (N odd), the slope of the PPC at θ_c is exactly given by the following formula (Tannoia, 2011):

$$P'(\theta_c) = \frac{N}{2^{N+1}} \binom{(N-1)}{\frac{N-1}{2}} \tag{7}$$

Since all items have the same difficulty, the score probability has a binomial distribution; then, the PPC is very well approximated by a *normal ogive* having at the critical ability the same slope as the PPC, i.e. with $\mu = \theta_c$ and $\sigma = \frac{1}{\sqrt{2\pi}P'(\theta_c)}$. Therefore, the critical misclassification interval for an ideal

test with N items of equal difficulty coincident with the mastery level, for N odd, can be estimated to be

$$\theta_c \pm \frac{2^{N+1}\text{probit}(\alpha)}{\sqrt{2\pi}N\binom{N-1}{\frac{N-1}{2}}} \tag{8}$$

where probit(x) denotes the *standard normal quantile*, i.e., the inverse of the standard normal cumulative distribution. For even N, an analogous derivation would be much more complicated, but an equally accurate estimate can be obtained by interpolation.

Using Stirling formula, one can see that for large N the critical amplitude is asymptotically proportional to $\frac{1}{\sqrt{N}}$, i.e. to the standard measurement error at θ_c. Empirically, it turns out that this holds also for non-ideal tests, provided the probability level α is chosen such that $MPC(\theta_0) > \alpha$, i.e., provided θ_0 is contained in the critical interval. This is confirmed in all computations made by the authors, as well as the fact that the PPC slope at θ_c is approximately equal to $\sqrt{\frac{F(\theta_c)}{2\pi}}$, $F(\theta_c)$ being the value of the test information function at θ_c. Even in the absence of a formal proof, it can thus be said that for a generic test the critical misclassification interval (for $\alpha \ll 0.5$) is approximately given by

$$\theta_c \pm \frac{\text{probit}(\alpha)}{\sqrt{I(\theta_c)}} \tag{9}$$

(for computational ease, probit(α) can be approximated, up to a factor, by the logistic cumulative function: probit$(\alpha) \approx \frac{\text{logit}(\alpha)}{1.7}$).

The critical interval for $\alpha = 0.5$, instead, is found in a different way (as already explained), and thus deserves a separate name:

Definition 6 *The* supercritical interval *is the ability range, bounded by the critical ability θ_c and the mastery level θ_0, where misclassification probability exceeds 0.5. It is always contained in the interval between $\Theta(s_0 - 1)$ and $\Theta(s_0)$ (by Prop. 3), and vanishes only if $\theta_c = \theta_0$.*

5 Generalisation to 2PL model

So far, the discussion has been restricted to the Rasch model; it is legitimate to ask to which extent it holds true if the item responses are assumed to be described by a more general IRT model. In particular, the 2PL model is commonly regarded as a more realistic description of the response process. The probability of correct response to the i-th item is then given by the formula

$$P_i(\theta) = \frac{1}{1 + e^{\alpha_i(\beta_i - \theta)}} \tag{10}$$

where α_i is the *discrimination parameter*.

It is known that in the 2PL model the number-right raw score is no longer a sufficient statistic for the examinee's ability; nevertheless, in the literature the TCC curve (still defined by $TCC(\theta) = \sum_i P_i(\theta)$, which gives the expectation value of the score for each ability θ) is used, for instance, for test equating purposes (Baker, 1992). In common practice, the pass criterion

in mastery tests is often given by a number-right score, no matter which IRT model is assumed: this happens, *a fortiori,* whenever a cut score is fixed without reference to the ability scale (situation A in the previous sections).

Now, *as long as number-right scoring is used,* the sources of misclassification remain the same as described in the previous sections, even if the item response probability is given by (10). There are still only $N-1$ observable abilities, the PPC can be computed as in the Rasch case, only using the 2PL probability (10) for the correct response. If the item discrimination parameters α_i is lower than 1 for most items, then the PPC has a lower slope at the critical ability w.r. to a Rasch test with the same difficulty spectrum, which entails a larger amplitude of the critical interval, in accordance with Birnbaum's statement that in an optimal test set the discrimination parameters should be the highest possible for all items.

However, assuming in the 2PL model a one-to-one correspondence between number-right scores and estimated abilities is incorrect. In fact, it is easy to prove (Baker, 1992) that the maximum likelihood estimate for the examinee's ability for a given response vector $\{u_i\}$ (whereby $u_i = 1$ if the answer to the i-th item is correct, and $u_i = 0$ otherwise) corresponds to the ability θ such that

$$\sum_i \alpha_i u_i = \sum_i \alpha_i P_i(\theta). \tag{11}$$

Hence, the appropriate score-ability correspondence is given by equating the *weighted score* on the l.h.s. of (11) to the *weighted TCC* given on the r.h.s.

The weighted score will be different for each response pattern: generically (namely, if the discrimination parameters α_i are all different, and if the sum of them over any item subset never coincides with the sum over a different subset) the weighted score can assume 2^N different values, in contrast with the number-right score which can take only $N+1$ values. For a given mastery ability θ_0, the (weighted) cut score still corresponds to the nearest observable ability on the right of θ_0. If the variance within the α_i is very little, the observable abilities will cluster around $N+1$ values; but if the discrimination variance is large enough, the observable abilities will instead tightly fill a large portion of the ability scale. In this case, the critical ability θ_c will always be very close to θ_0 and the supercritical interval will be negligible.

To illustrate the situation, assume for the items of the test set (1) the following discrimination parameters (instead of $\alpha_i \equiv 1$ as in sect. 2):

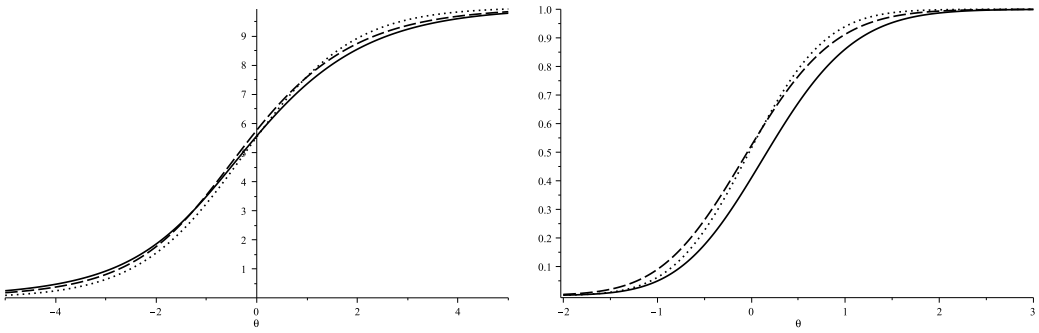
$$\begin{array}{lllll} \alpha_1 = 1.30 & \alpha_2 = 1.09 & \alpha_3 = 0.46 & \alpha_4 = 1.21 & \alpha_5 = 0.66 \\ \alpha_6 = 0.86 & \alpha_7 = 0.89 & \alpha_8 = 0.79 & \alpha_9 = 0.90 & \alpha_{10} = 0.71 \end{array} \tag{12}$$

The average discrimination parameter is $\bar{\alpha} = 0.89$. The mastery level is now assumed to be $\theta_0 = 0.20$; the following plots allow to compare the graphs for the TCC, the PPC and the MPC obtained, respectively, for

1. (solid curves): the 2PL test set with discrimination parameters (12), upon weighted scoring, with a (weighted) cut score of 6.19 (NOTE: to allow comparison with number-right scores, all weighted scores have been linearly rescaled – dividing them by $\bar{\alpha}$ – so that the maximum score is always N=10);

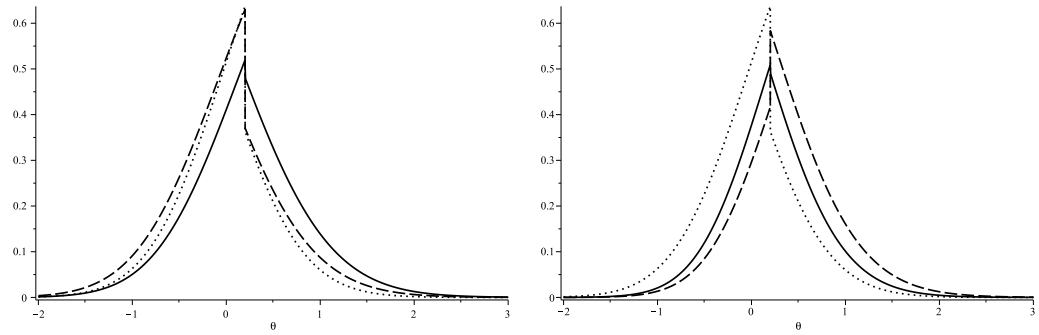
2. (dashed curves): the same 2PL test set, but assuming ordinary number-right scoring (the proper cut score turns out to be 6)

3. (dotted curves): a test set with the same difficulties (1) but $\alpha_i \equiv 1$ (Rasch case); for this test set the number-right and the weighted scores coincide, and the cut score is also 6 (the MPC for this case is the same as in Fig. 2c).



(a) TCC **(b)** PPC

Figure 7: Comparison of TCC and PPC for the 2PL test with parameters given by (1) and (12), under weighted scoring (solid) and number-right scoring (dashed). Dotted curves refer to a test set with the same difficulties β_i but $\alpha_i \equiv 1$.



(a) MPC for the 2PL test set of Fig. 7 **(b)** MPC for another 2PL test set

Figure 8: Comparison of Misclassification Probability Curves

The figures show that the TCC is very close for the three cases. As long as

number-right scoring is used, in this example the effect of the α_i distribution on the Pass Probability Curve only amounts to a lower slope (due to $\bar{\alpha} < 1$); a noticeable shift of the PCC occurs instead if weighted scoring is used. The truly relevant differences, however, can be seen in the MPC (Fig. 8.a): under weighted scoring, the critical ability becomes close to θ_0, and the MPC discontinuity almost disappears. In this case, the MPC curve would be well approximated by Rudner's method. This does not mean that the overall misclassification rate would decrease for any population: weighted scoring yields, in this example, a lower amount of false masters but a larger amount of false non-masters.

The inspection of MPC plots allows to answer some relevant questions, that it is important to keep distinct. If the variance of the discrimination parameters of the items is significant, then *adopting the appropriate weighted scoring* produces a misclassification probability curve which is much closer to Rudner's curve. To this purpose, it is the *scoring formula* that matters, not just the use of the 2PL probabilities (10). The result of the test (pass or fail) will be different for some examinees while using a different scoring formula: therefore, it is the actual reliability of the test – not only its estimate – which changes (yet, the overall reliability will not necessarily be higher under weighted scoring).

A separate question is the following: if number-right scoring is used, to which extent ignoring the variation of the discriminating parameter among items, and thus applying the Rasch model to compute misclassification prob-

abilities, would lead to an incorrect assessment of the classification consistency? The difference between the dotted and the dashed curve in Fig. 8.a is little and is actually due to the fact that the *average* discrimination parameter of the test set (12) is lower than 1. But it is not always true that the MPC for a 2PL test, under number-right scoring, is close to the MPC of a Rasch test with the same difficulty spectrum. With a different simulated distribution of the discrimination parameters (with $\bar{\alpha} = 1.08$, still leaving difficulties unchanged), the (unweighted) cut score for $\theta_0 = 0.20$ jumps to 7, and the MPCs become as in Fig. 8.b: the supercritical interval shifts to the right of the mastery ability. The large difference between the dotted and the dashed curve is due to the discontinuous dependence of the cut score on the item parameters. Such a discrepancy, therefore, may occur when the TCC score corresponding to the mastery ability is close to an integer value (a fact that could be detected already in the Rasch setup).

Weighted scoring does reduce such instabilities, and might therefore seem to yield more robust results, but actually raises a different problem. Weighted scoring rests on the knowledge of the α_i: in the discussion it has been assumed so far that the discrimination parameters were exactly known. Any uncertainty on the values of the α_i is reflected in a *score indeterminacy,* a situation never met in number-right scoring (which is insensitive to item parameters): such indeterminacy is a new potential source of misclassification. To judge its effect size, assume a measurement error of ± 0.1 in the α_i listed in (12): numerical computation shows that the resulting score standard de-

viation (close to the cut score) corresponds to an uncertainty of ± 0.09 logit in the ability scale. This is an *additional* uncertainty, which has nothing to do with the standard error of the ability estimate (related to the information function value): the latter is indeed much larger, but arises from the probabilistic nature of the response process, which is exactly represented by the MPC. The new uncertainty concerns instead the score to be assigned (by the scoring formula) to a given response pattern, and fails to be rendered in the MPC. In our example, the width of this "score indeterminacy interval" is comparable to the width of the supercritical interval that would be found with number-right scoring. In other terms, the "improvement" of the theoretical misclassification curve obtained by adopting weighted scoring in the 2PL model may be somehow fictitious if the discrimination parameters are not known with good accuracy, for another source of misclassification arises.

6 Conclusions

It has been shown that the *intrinsic* reliability of a test set is described by the Misclassification Probability Curve. The *overall reliability for a given population* (as measured by the total misclassification incidence) will depend on the test's MPC *and* on the population ability distribution.

A comparison with Birnbaum's setting of the reliability problem (Lord & Novick, 1968) helps in focusing the picure emerging from the present discussion. Birnbaum took as starting point a predetermined *indifference interval*,

bounded by two reference abilities ("definitely low" and "definitely high"). Given this interval, Birnbaum extensively discussed (without restricting to a specific item response model) *which scoring rule* and *which cut score* would minimize the misclassification probability. Here, instead, the starting point is either the cut score (situation A) or the mastery level (situation B), and ability intervals related to misclassification issues are derived as a result.

Situation A is relatively simple: if one resists the temptation to identify the mastery level with $\Theta(s_0)$ (the estimated ability corresponding to the cut score), and takes instead the critical ability θ_c, the misclassification probability cannot exceed 0.5 for any latent ability. Then, classification accuracy can be improved by raising the slope of the Pass Probability Curve at θ_c. One can expect to get this result by increasing the number of items and/or by choosing items with difficulties closer to θ_c. Notice, however, that any change of the test set alters the value of θ_c itself; moreover, if the cut score was decided after some process of item appraisal, this has to be redone if the test set is modified. What can be safely said is that among different test sets (each with its own cut score) *yielding the same critical ability, and therefore assessing the same mastery level*, the most reliable is the one which has the highest PPC slope (or, equivalently, the highest value of the information function at θ_c).

A test can be regarded as *definitely reliable* outside the critical interval (for a suitably low value of α, such as the standard values $\alpha = 0.1$ or $\alpha = 0.05$). The amplitude of the latter is related by (9) to the value of the test

information function at θ_c. The boundary points of the critical interval might thus be regarded as "definitely low" and "definitely high" abilities (relative to the confidence level α): yet, there is a conceptual difference with respect to Birnbaum's approach. The critical interval is not an "indifference interval": classification is performed, and misclassification is considered, also within that interval.

In situation B, the mastery level θ_0 being preassigned, the critical ability θ_c will often not coincide with it. Then, *two* ability intervals should be considered in connection to pass-fail reliability. The critical interval still encodes the reliability of the test *far from the mastery level*. Upon narrowing the critical interval, a larger portion of the population will fall in the "safe" region. However, misclassification is much more likely to occur for abilities close to the mastery level: in situation B the supercritical interval may become the primary locus of misclassification. The width of this interval is unrelated to the test information function: for instance, if one takes an "optimal" test set with an odd number of items, and *adds* a further item with difficulty equal to θ_0, this will actually *enlarge* the supercritical interval, thus reducing the reliability around the threshold, although the information function is increased.

In conclusion, the critical and the supercritical interval, together, allow to foresee the overall functioning of the mastery test for a given population, and to judge whether reliability issues should be confronted by increasing the test information function, or rather by trying to match the critical ability of

the test with the sought mastery level.

Which test design would be optimal actually depends on the population, and the possibly different "costs" of false masters and false non-masters should also be taken into account. There is no simple recipe to single out the most reliable test set from a finite pool of calibrated items. Computing for a test set the critical ability θ_c and the value of the test information function at θ_c is quite viable, and allows to compare a number of different test sets and select the one having both the narrowest critical and the narrowest supercritical interval: this is likely to be the most reliable test set in the group. To adjust for asymmetries in the population distribution with respect to the mastery level, the expected misclassification rates should be computed and compared as well, provided a reliable estimate of the ability distribution in the population is available.

If the 2PL model is considered instead of the Rasch model, the picture remains similar as long as unweighted scoring is used. If, instead, the proper (weighted) scoring is implemented and a weighted cut score is fixed accordingly, the distance between consecutive observable abilities is drastically reduced and the supercritical interval becomes negligible. However, in this case a different source of misclassification in the vicinity of the cut score may arise – the uncertainty in the score to be assigned to each response pattern – unless the item discrimination parameters are known with high accuracy: this requires item calibration on a larger population (for this reason we did not consider here the 3PL model, where the accuracy of item parameter

estimates is a delicate issue).

References

Baker, F.B. 1992. *Item Response Theory: Parameter estimation techniques*. Statistics: textbooks and monographs. Marcel Dekker.

Gatti, G.G., & Buckendahl, C.W. 2006. On Correctly Classifying Examinees. *In: Annual meeting of the American Educational Research Association, San Francisco CA, April 2006*.

Guo, F. 2006. Expected Classification Accuracy Using Latent Distributions. *GMAC Research Reports*.

Huynh, H. 1980. Statistical inference for false positive and false negative error rates in mastery testing. *Psychometrika*, **45**, 107–120.

Huynh, H. 1982. Assessing Efficiency of Decisions in Mastery Testing. *Journal of Educational Statistics*, **7**(1), 47–63.

Huynh, H. 1990. Computation and Statistical Inference for Decision Consistency Index Based on the Rasch Model. *Journal of Educational Statistics*, **15**(4), 353–368.

Khidr, A.M., & Abdelnasser, M.T. 1982. Decomposing the sum of independent Bernoulli variates to its components. *Indian J. pure appl. Math.*, **49**, 223–245.

Livingston, S.A., & Lewis, Ch. 1995. Estimating the Consistency and Accuracy of Classifications Based on Test Scores. *Journal of Educational Measurement*, **32**(2), 179–197.

Lord, F.M., & Novick, M.R. 1968. *Statistical Theories of Mental Test Scores*. Addison-Wesley. (contributed by A. Birnbaum). Chap. 17 and 19.

Lord, F.M., & Wingersky, M.S. 1984. Comparison of IRT true-score and equipercentile observed-score equatings. *Applied Psychological Measurement*, **8**, 453–461.

Rudner, L.M. 2005. Expected Classification Accuracy. *Practical Assessment, Research & Evaluation*, **10**(13).

Tannoia, C. 2011. *Pass-fail reliability for multiple-choice tests with cut scores*. MSc Thesis, University of Turin.

van der Linden, W.J. 1998. A Decision Theory Model for Course Placement. *Journal of Educational and Behavioral Statistics*, **23**(1), 18–34.

Wainer, H., Wang, X.A., & Bradlow, E.T. 2005. A Bayesian method for evaluating passing scores: The PPoP curve. *Journal of Educational Measurement*, **42**(3), 271–281.

Wilcox, R.R. 1977. Estimating the likelihood of false-positive and false-negative decisions in mastery testing: an empirical Bayes approach. *Journal of Educational Statistics*, **2**, 289–307.

Young, M.J., & Yoon, B. 1998. Estimating the Consistency and Accuracy of Classifications in a Standards-Referenced Assessment. *CSE Technical Report*.